Shifts in Rating Bias due to Scale Saturation

Kanika Kalra, Manasi Patwardhan, Shirish Karande
TCS Research, 54-B Hadpasar Industrial Estate, Pune 411013, India
kalra.kanika@tcs.com, manasi.patwardhan@tcs.com, shirish.karande@tcs.com

Abstract

Rating scales are used widely in online surveys and cardinal peer evaluations. There has been work on modelling bias and reliability of workers under the assumptions that these parameters remain static and unaffected by the observed distribution of the underlying tasks. Anecdotal experiences in grading often points to the contrary and consequently in this work-in-progress we seek to construct experiments that clearly demonstrate bias changes. Better understanding of this phenomenon can lead to improved models and algorithms. We specifically observed that the order of tasks can lead to scale saturation which may lead to bias shifts.

Introduction

In crowdsourcing setup the input provided by a worker to every assigned task is considered to be independent. Many task allocation, crowd consensus and pricing algorithms are built on this base assumption. In this paper we have conducted an experiment to demonstrate that, for certain type of tasks (rating objects in this case), an input provided by a worker to a task is a function of the distribution of a signal carried by the task instances, observed prior to that instance, in a specific order. For example, in the task of rating the amount of damage observed in a car image; the input provided by a worker to the n^{th} image is dependent on the distribution of intensity of damage observed until that point.

Till date many approaches have studied distinct types of biases that play a role in crowd-setup. Task based biases include bias caused sue to task perplexity (Kamar, Kapoor, and Horvitz 2015), visual similarity of distinct tasks (Meta 2016) and placement of task on the worker interface (Kaufmann, Schulze, and Veit 2011). Apart from task biases, the literature discusses distinct types of worker biases. Social (Antin and Shaw 2012), cultural (Kaufmann, Schulze, and Veit 2011) linguistics or gender (Otterbacher 2015) biases are caused for the workers with distinct demographics. Complexity based task ordering affects worker positively improving their efficiency (Cai, Iqbal, and Teevan 2016). A peer bias affects peer grading (Piech et al. 2013). A bias towards prior responses to a task gets introduced when a worker is exposed to such responses (Chatterjee, Mukhopadhyay, and Bhattacharyya 2016). Workers may get biased towards their own responses to a set of tasks (Faltings et al. 2014). For example, for a task of grammatical error correction, where most sentences have no errors, then the errors in the next immediate set of task instances may not get noticed.

Figure 1: Image Sequence: The categorization as per the perception of the authors

In this paper, we are neither modeling task bias nor worker bias; whereas we sought to create a rating task with a task sequence that could demonstrate a bias shift of a worker. Through an experiment we try to demonstrate that this shift in bias is due to the recalibration effect that occurs as a result of the observed distribution of the underlying tasks making the worker reach a point where extreme rating values are exhausted. We henceforth call this point as a saturation point. It has been discussed in literature that descriptiveness and rules for rating can help neutralize the bias of a worker. We therefore specifically did neither; however, it was observed that workers would seek to form their own rules for rating, which tend to change the rating scale of the worker after the saturation point.

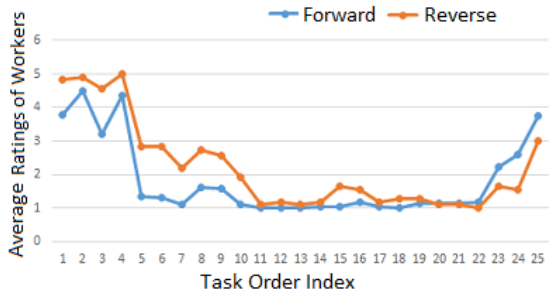

Figure 2: Image Ratings for Forward and Reverse Order

Experimental Design and Observations

As a part of the experiment we display distinct types of objects to workers and the task is to rate the objects on the scale of 1 to 5, 1 showcasing weak and 5 showcasing strong signal. Distinct objects that we have taken into consideration are: (i) Translation for translation quality (ii) Images of a monument for aesthetics quality (perception of beauty) (iii) Images of streets for safety (iv) Images of cars for damage.

The initial set of experiments performed have led to the following set of observations. Objects selection plays a very critical role in the successful design of the experiment. For translation and aesthetics oriented tasks, the worker with higher expertise showcase lower shift in bias. Also for the translation task, the variations in diction made it hard to control the experiment. For translation and street image safety task every source sentence or a street image is distinctively different than the earlier instances. This leads to interruption in the flow of work breaking the continuity of the worker. This allows worker to refresh his bias for every new instance of task he sees. This points us to a fact that we need to choose a task which would maintain a flow in the sequence avoiding discontinuity. We further observe that the multidimensional crowd wisdom made it hard to construct a task sequence which would illicit bias shift. For example, an Indian worker, who does not have any context for the street images of Newyork city, may not share the perception of safety with an American worker. Finally, we decide to choose the task of rating car damage, this task seemed to have a large consensus when piloted for pair-wise ranking, proving it to be more objective as compared to the other tasks.

The order in which the object instances appear is very crucial. This order is hand-crafted to lead to scale saturation. Figure 1 displays the sequence of 25 images, shown in our experiment to distinct set of workers, in the forward as well as reverse order. The sequence is designed to have its saturation point nearly at index 15 in the forward order. Our crowd consists of 44 volunteers, who we know are not spammers, but they are unaware of the experiment design and the motivation for task sequence. We have ensured that a worker provides rating for all 25 instances of objects in one go without any break to maintain the continuity of the task.

Results

Figure 2 illustrates the average ratings of all the workers for each image in the sequence. The comparison of the ratings in

Figure 3: KL Divergence of Rating Distributions of Forward and Reverse Order

forward and reverse order reveals the bias shift. In particular the saturation region $\approx (11-22)$ is a point of inflexion, where on either side of the saturation one can observe an upward shift in damage estimates. This upward shift is due to the recalibration made by workers in the saturation region.

We further observe the KL-Divergence graph (Figure 3), which demonstrates the deviation in the rating distributions of the same task, but received during the forward and reverse order. The less divergence at the saturation region, clearly indicates that many workers are forced to give same ratings at the saturation region, irrespective of the sequence order.

Discussions

We propose a plausible model that formulates the shift in bias on the similar lines of (Raman and Joachims 2014), which defines a model for peer grading having a certain worker bias and reliability.

$$y_t^{(w)} \sim \mathcal{N}(s_t + b_w + I(sat) * b_{wsat}, 1/\eta_w) \qquad (1)$$

In the above equation $y_t^{(w)}$ is the response provided by a worker w to task t, which is drawn from a Gaussian distribution. s_t is the actual label for that task, b_w is the inherent bias of the worker, $I(sat) \in 0, 1$, 1 when the saturation point is reached, b_{wsat} is the shift in the bias and η_w is the worker reliability. The saturation point can be identified by observing

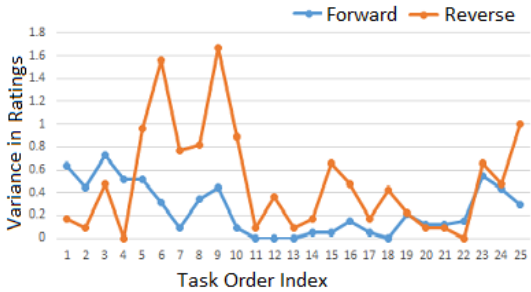

Figure 4: Rating Variances for Forward and Reverse Order

the variance of the worker ratings as the sequence progresses and identifying a point where variance converges to 0 (Figure 4). This being our initial attempt, more elegant models can be formulated for the phenomenon.

References

Antin, J., and Shaw, A. 2012. Social desirability bias and self-reports of motivation: a study of amazon mechanical turk in the us and india. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2925–2934. ACM.

Cai, C. J.; Iqbal, S. T.; and Teevan, J. 2016. Chain reactions: The impact of order on microtask chains. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, 3143–3154. ACM.

Chatterjee, S.; Mukhopadhyay, A.; and Bhattacharyya, M. 2016. Consensus of dependent opinions. *arXiv preprint arXiv:1609.01408*.

Faltings, B.; Jurca, R.; Pu, P.; and Tran, B. D. 2014. Incentives to counter bias in human computation. In *Second AAAI conference on human computation and crowdsourcing*.

Kamar, E.; Kapoor, A.; and Horvitz, E. 2015. Identifying and accounting for task-dependent bias in crowdsourcing. In *Third AAAI Conference on Human Computation and Crowdsourcing*.

Kaufmann, N.; Schulze, T.; and Veit, D. 2011. More than fun and money. worker motivation in crowdsourcing-a study on mechanical turk. In *AMCIS*, volume 11, 1–11.

Meta, E. 2016. *Question Bias in Repetitive Crowdsourcing Tasks*. Ph.D. Dissertation.

Otterbacher, J. 2015. Crowdsourcing stereotypes: Linguistic bias in metadata generated via gwap. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, 1955–1964. ACM.

Piech, C.; Huang, J.; Chen, Z.; Do, C.; Ng, A.; and Koller, D. 2013. Tuned models of peer assessment in moocs. *arXiv preprint arXiv:1307.2579*.

Raman, K., and Joachims, T. 2014. Methods for ordinal peer grading. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, 1037–1046. ACM.